Multiset Combinatorial Batch Codes

Hui Zhang[1], **Eitan Yaakobi**[1], and **Natalia Silberstein**[2]

[1]Computer Science Department, Technion—Israel Institute of Technology, Haifa, Israel
[2]Yahoo! Labs, Haifa, Israel
emails: *huizhang@cs.technion.ac.il, yaakobi@cs.technion.ac.il, natalys@cs.technion.ac.il*

***Abstract*—Batch codes, first introduced by Ishai, Kushilevitz, Ostrovsky, and Sahai, mimic a distributed storage of a set of n data items on m servers, in such a way that any batch of k data items can be retrieved by reading at most some t symbols from each server. *Combinatorial batch codes*, are replication-based batch codes in which each server stores a subset of the data items.**

In this paper, we propose a generalization of combinatorial batch codes, called *multiset combinatorial batch codes* (*MCBCs*), in which n data items are stored in m servers, such that any multiset request of k items, where any item is requested at most r times, can be retrieved by reading at most t items from each server. The setup of this new family of codes is motivated by recent work on codes which enable high availability and parallel reads in distributed storage systems. The main problem under this paradigm is to minimize the number of items stored in the servers, given the values of n, m, k, r, t, which is denoted by $N(n, k, m, t; r)$. We first give a necessary and sufficient condition for the existence of MCBCs. Then, we present several bounds on $N(n, k, m, t; r)$ and constructions of MCBCs. In particular, we determine the value of $N(n, k, m, 1; r)$ for any $n \geqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1} - (m-k+1)A(m, 4, k-2)$, where $A(m, 4, k-2)$ is the maximum size of a binary constant weight code of length m, distance four and weight $k-2$. We also determine the exact value of $N(n, k, m, 1; r)$ when $r \in \{k, k-1\}$ or $k = m$.

I. Introduction

Batch codes were first introduced by Ishai *et al.* in [11] as a method to represent the distributed storage of a set of n data items on m servers. These codes were originally motivated by several applications such as load balancing in distributed storage, private information retrieval, and cryptographic protocols. Formally, these codes are defined as follows [11].

Definition 1.

1) *An* (n, N, k, m, t) batch code *over an alphabet* Σ, *encodes a string* $x \in \Sigma^n$ *into an* m*-tuple of strings* $y_1, \ldots, y_m \in \Sigma^*$ *(called* buckets *or* servers*) of total length* N, *such that for each* k*-tuple (called* batch *or* request*) of distinct indices* $i_1, \ldots, i_k \in [n]$, *the* k *data items* $x_{i_1}, \ldots, x_{i_k}$ *can be decoded by reading at most* t *symbols from each server.*
2) *An* (n, N, k, m, t) multiset batch code *is an* (n, N, k, m, t) *batch code which also satisfies the following property: For any multiset request of* k *indices* $i_1, \ldots, i_k \in [n]$ *there is a partition of the buckets into* k *subsets* $S_1, \ldots, S_k \subseteq [m]$ *such that each item* x_{i_j}, $j \in [k]$, *can be retrieved by reading at most* t *symbols from each bucket in* S_j.

Yet another class of codes, called *combinatorial batch codes* (CBC), is a special type of batch codes in which all encoded symbols are copies of the input items, i.e., these codes are replication-based. Several works have considered codes under this setup; see e.g. [1]–[8], [12], [14], [15]. However, note that combinatorial batch codes are not multiset batch codes and don't allow to request an item more than once.

Motivated by the works on codes which enable parallel reads for different users in distributed storage systems, for example, codes with locality and availability [13], [16], we introduce a generalization of CBCs, named *multiset combinatorial batch codes*.

Definition 2. *An* $(n, N, k, m, t; r)$ multiset combinatorial batch code (MCBC) *is a collection of subsets of* $[n]$, $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ *(called* servers*) where* $N = \sum_{j=1}^{m} |C_j|$, *such that for each multiset request* $\{i_1, i_2, \ldots, i_k\}$, *in which every element in* $[n]$ *has multiplicity at most* r, *there exist subsets* $D_1, \ldots, D_m$, *where for all* $j \in [m]$, $D_j \subseteq C_j$ *with* $|D_j| \leqslant t$, *and the multiset union*[1] *of* D_j *for* $j \in [m]$ *contains the multiset request* $\{i_1, i_2, \ldots, i_k\}$.

In other words, an $(n, N, k, m, t; r)$-MCBC is a coding scheme which encodes n items into m servers, with total storage of N items, such that any multiset request of items of size at most k, where any item can be repeated at most r times, can be retrieved by reading at most t items from each server. In particular, when $r = 1$ we obtain a combinatorial batch code, and when $r = k$ and $t = 1$ we obtain a multiset batch code based on replication. When $t \neq 1$, retrieval of a multiset for the MCBCs does not essentially partition the set of servers as for multiset batch codes.

Example 1. Let us consider the following $(n = 5, N = 15, k = 5, m = 5, t = 1; r = 2)$ MCBC,

1	1	2	2	3
3	4	3	4	4
5	5	5	5	5

where the i-th column contains the indices of items stored in the server $C_i \in \mathcal{C}$, $i \in [5]$. It is possible to verify that the code $\mathcal{C}$ satisfies the requirements of a $(5, 15, 5, 5, 1; 2)$-MCBC. For example, the multiset request $\{3, 3, 4, 4, 5\}$ can be read by taking the subsets $D_1 = \{3\}$, $D_2 = \{4\}$, $D_3 = \{3\}$, $D_4 = \{4\}$, $D_5 = \{5\}$. □

Similarly to the original problem of combinatorial batch codes, the goal in this paper is to minimize the total storage N given the parameters n, m, k, t and r of an MCBC. Let $N(n, k, m, t; r)$ be the smallest N such that an

[1]For any $i \in [n]$, the multiplicity of i in the multiset union of the sets D_j for $j \in [m]$ is the number of subsets that contain i, that is $|\{j \in [m] : i \in D_j\}|$.

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

$(n, N, k, m, t; r)$-MCBC exists. An MCBC is called *optimal* if N is minimal given n, m, k, t, r. In this paper, we focus on the case $t = 1$, and thus omit t from the notation and write it as an $(n, N, k, m; r)$-MCBC and its minimum storage by $N(n, k, m; r)$. In case $r = 1$, i.e. when an MCBC is a CBC, we further omit r and write it as an (n, N, k, m)-CBC and its minimum storage as $N(n, k, m)$.

For CBCs, a significant amount of work has been done to study the value $N(n, k, m)$, and the exact value has been determined for a large range of parameters. For a list of the known results we refer the reader to [2], [3], [5], [6], [12], [15].

The rest of the paper is organized as follows. In Section II, we give a necessary and sufficient condition for the existence of MCBCs. In Section III, we give several bounds on MCBCs. In Section IV, we present constructions of MCBCs. Lastly, Section V concludes the paper. Due to the lack of space, some of the proofs in the paper are deferred to the full version.

II. Set Systems and the Multiset Hall's Condition

A *set system* is a pair $(V, \mathcal{C})$, where V is a finite set of *points* and $\mathcal{C}$ is a collection of subsets of V (called *blocks*). Given a set system $(V, \mathcal{C})$ with a points set $V = \{v_1, v_2, \ldots, v_n\}$ and a blocks set $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, its *incidence matrix* is an $m \times n$ matrix M, given by

$$M_{i,j} = \begin{cases} 1 & \text{if } v_j \in C_i, \\ 0 & \text{if } v_j \notin C_i. \end{cases}$$

If M is the incidence matrix of the set system $(V, \mathcal{C})$, then the set system having incidence matrix $M^\top$ is called the *dual set system* of $(V, \mathcal{C})$.

Let $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ be an $(n, N, k, m; r)$-MCBC. Similarly to the study of CBCs, by setting $V = [n]$, we consider the set system $(V, \mathcal{C})$ of the MCBC. In addition, we denote the set system $(X, \mathcal{B})$ which is given by $X = [m]$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ where for each $i \in [n]$, $B_i \subseteq X$ consists of the servers that store the i-th item. Then, it is readily verified that $(X, \mathcal{B})$ is the dual set system of $(V, \mathcal{C})$. We note that a set system $(V, \mathcal{C})$ of this form or its dual set system $(X, \mathcal{B})$ uniquely determines an MCBC and thus in the rest of the paper we will usually refer to an MCBC by its set system or its dual set system.

Example 2. In Table I, we give a $(20, 80, 16, 16)$-CBC from [15] based on an *affine plane* of order 4. Here, $V = [20]$, each column contains the indices of items stored in a server $C_i \in \mathcal{C}$ and also forms a block of the set system $(V, \mathcal{C})$. Here, the dual set system is $X = [16]$ and $\mathcal{B} =$

$$\{\{1,5,9,13\},\{2,6,10,14\},\{3,7,11,15\},\{4,8,12,16\},\\ \{1,6,11,16\}\{2,5,12,15\},\{3,8,9,14\},\{4,7,10,13\},\\ \{1,8,10,15\},\{2,7,9,16\},\{3,6,12,13\},\{4,5,11,14\},\\ \{1,7,12,14\},\{2,8,11,13\},\{3,5,10,16\},\{4,6,9,15\},\\ \{1,2,3,4\},\{5,6,7,8\},\{9,10,11,12\},\{13,14,15,16\}\}.$$

The incidence matrix of the CBC given above is as follows, where the indices of nonzero entries in the i-th row, $i \in [16]$, correspond to the indices of items stored in the i-th server C_i.

$$\begin{pmatrix}
1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&1&0&0&0\\
0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&1&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&1&0&0&0\\
1&0&0&0&0&1&0&0&0&0&0&1&0&0&1&0&0&1&0&0\\
0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0\\
0&0&1&0&0&0&0&1&0&1&0&0&1&0&0&0&0&1&0&0\\
0&0&0&1&0&0&1&0&1&0&0&0&0&1&0&0&0&1&0&0\\
1&0&0&0&0&0&1&0&0&1&0&0&0&0&0&1&0&0&1&0\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0&1&0&0&0&1&0\\
0&0&1&0&1&0&0&0&0&0&0&1&0&1&0&0&0&0&1&0\\
0&0&0&1&0&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0\\
1&0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&1\\
0&1&0&0&0&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1\\
0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&1\\
0&0&0&1&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1
\end{pmatrix}$$

□

In the rest of this section we let $(V, \mathcal{C})$ with $V = [n]$ and $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ be a set system, and $(X, \mathcal{B})$ with $X = [m]$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ be its dual set system. The following theorem states a necessary and sufficient condition on the dual set system to form a construction of CBCs.

Theorem 3. *[[12]] The set system $(V, \mathcal{C})$ is an (n, N, k, m)-CBC if and only if its dual set system $(X, \mathcal{B})$ satisfies the following* Hall's condition*:*
for all $h \in [k]$, and any h distinct blocks $B_{i_1}, B_{i_2}, \ldots, B_{i_h} \in \mathcal{B}$, $|\cup_{j=1}^{h} B_{i_j}| \geqslant h$.

In this paper, we present a generalization of the Hall's condition, named the *multiset Hall's condition*, and provide a necessary and sufficient condition for the construction of MCBCs.

Theorem 4. *The set system $(V, \mathcal{C})$ is an $(n, N, k, m; r)$-MCBC if and only if its dual set system $(X, \mathcal{B})$ satisfies the following* multiset Hall's condition*:*
for all $h \in [\lceil \frac{k}{r} \rceil]$, and any h distinct blocks $B_{i_1}, B_{i_2}, \ldots, B_{i_h} \in \mathcal{B}$, $|\cup_{j=1}^{h} B_{i_j}| \geqslant \min\{hr, k\}$.

Proof: ($\Rightarrow$) Assume that $(V, \mathcal{C})$ is an $(n, N, k, m; r)$-MCBC, and let $i_1, i_2, \ldots, i_h \in V$ for some $h \in [\lceil \frac{k}{r} \rceil]$ be the indices of some h different items. Then, the set $\cup_{j \in [h]} B_{i_j}$ corresponds to the indices of all the servers that contain these items.

If $h \leqslant \lfloor \frac{k}{r} \rfloor$, let us consider the multiset request $\{i_1, \ldots, i_1, i_2, \ldots, i_2, \ldots, i_h, \ldots, i_h\}$ where each of the h elements is requested r times. Since it is possible to read from each server at most one item, the number of servers that contain these h items has to be at least hr, that is $|\cup_{j \in [h]} B_{i_j}| \geqslant hr$. Similarly, if $h = \lceil \frac{k}{r} \rceil$, then we need k servers for the multiset request of size k on $i_1, i_2, \ldots, i_h$ where each i_j, for $j \in [h]$, is requested at most r times, and so $|\cup_{j \in [h]} B_{i_j}| \geqslant k$. Together we conclude that $|\cup_{j=1}^{h} B_{i_j}| \geqslant \min\{hr, k\}$.

($\Leftarrow$) We construct a new set system $(U, \mathcal{F})$ with $U = [rn]$ and $\mathcal{F} = \{F_1, F_2, \ldots, F_m\}$ where for $i \in [m]$, $F_i = \{c + jn : c \in C_i, j \in [0, r-1]\}$. Let $U^{(\ell)} = \{\ell + jn : j \in [0, r-1]\}$ for $\ell \in [n]$. We first show the following claim.

TABLE I
A $(20, 80, 16, 16)$-CBC

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
5	6	7	8	6	5	8	7	7	8	5	6	8	7	6	5
9	10	11	12	12	11	10	9	10	9	12	11	11	12	9	10
13	14	15	16	15	16	13	14	16	15	14	13	14	13	16	15
17	17	17	17	18	18	18	18	19	19	19	19	20	20	20	20

Claim 1 *$(V, \mathcal{C})$ is an $(n, N, k, m; r)$-MCBC if $(U, \mathcal{F})$ is an (rn, rN, k, m)-CBC.*

Proof: Suppose that $(U, \mathcal{F})$ is an (rn, rN, k, m)-CBC. For any multiset request Q of $(V, \mathcal{C})$ where each $\ell \in [n]$ appears r_ℓ times with $0 \leqslant r_\ell \leqslant r$ and $\sum_{\ell=1}^{n} r_\ell = k$, consider the request P of $(U, \mathcal{F})$ which contain any r_ℓ distinct elements in $U^{(\ell)}$. Since $(U, \mathcal{F})$ is an (rn, rN, k, m)-CBC, P can be read by taking $D_j \subseteq F_j$, $|D_j| \leqslant 1$ for $j \in [m]$. Then Q can be read from the servers $\{C_j : j \in [m], |D_j| = 1\}$. Therefore, $(V, \mathcal{C})$ is an $(n, N, k, m; r)$-MCBC. ∎

Let $(X = [m], \mathcal{G} = \{G_1, \ldots, G_{nr}\})$, be the dual set system of $(U, \mathcal{F})$, so G_i, for $i \in [rn]$, is the set of servers that contain the i-th item in $(U, \mathcal{F})$. We show that $(X, \mathcal{G})$ satisfies the Hall's condition. For any $i_1, i_2, \ldots, i_h \in [rn]$, $h \in [k]$, let r_ℓ denote the number of elements in $U^{(\ell)}$ for $\ell \in [n]$. Then $\left|\cup_{j=1}^{h} G_{i_j}\right| = \left|\cup_{\ell: r_\ell \neq 0} B_\ell\right|$.

Let $a = |\{\ell : r_\ell \neq 0\}|$. By the multiset Hall's condition, when $a \leqslant \lceil \frac{k}{r} \rceil$, $\left|\cup_{\ell: r_\ell \neq 0} B_\ell\right| \geqslant \min\{ar, k\}$; when $\lceil \frac{k}{r} \rceil < a \leqslant k$,

$$\left|\bigcup_{\ell: r_\ell \neq 0} B_\ell\right| \geqslant \min\{r \left\lceil \frac{k}{r} \right\rceil, k\} \geqslant k = \min\{ar, k\}.$$

Since $h = \sum_{\ell: r_\ell \neq 0} r_\ell \leqslant ar$ and $h \leqslant k$, we always have $|\cup_{j=1}^{h} G_{i_j}| \geqslant h$ for any $h \in [k]$, that is $(X, \mathcal{G})$ satisfies the Hall's condition. Hence, $(U, \mathcal{F})$ is an (rn, rN, k, m)-CBC by Theorem 3, and by Claim 1 $(V, \mathcal{C})$ is an $(n, N, k, m; r)$-MCBC. ∎

In the following, when constructing an MCBC, we always construct its dual set system $(X, \mathcal{B})$, and check if it satisfies the multiset Hall's condition from Theorem 4. By adding an asterisk, we let $(X, \mathcal{B})^*$ denote its dual set system $(V, \mathcal{C})$.

Example 3. By checking the multiset Hall's condition, it is possible to verify that Example 2 gives a construction of $(20, 80, k, 16; r)$-MCBC for any pair $(k, r) \in \{(16, 1), (11, 2), (10, 3), (7, 4)\}$. □

In the following sections, we will give several bounds and constructions of MCBCs.

III. Bounds of MCBCs

In this section, we give several bounds of MCBCs. We first state a lemma with some basic properties on the value of $N(n, k, m; r)$.

Lemma 5.

(i) $N(n, k, m; r) \geqslant rn$.
(ii) $N(n, k, m; r) \geqslant N(n, k, m; i)$ *for* $i \in [r-1]$.
(iii) $N(n, k, m; k) = kn$.
(iv) $\frac{1}{r} N(nr, k, m) \leqslant N(n, k, m; r) \leqslant N(rn, k, m)$.
(v) $N(n, k, m; r) \leqslant rN(n, \left\lceil \frac{k}{r} \right\rceil, \lfloor \frac{m}{r} \rfloor)$.

Let $(V, \mathcal{C})$ be a set system of an $(n, N, k, m; r)$-MCBC and let $(X, \mathcal{B})$ be its dual set system. For $i \geqslant 0$, we denote by A_i the number of subsets in $\mathcal{B}$ of size i. Note that for $i < r$, $A_i = 0$ since every item is contained in at least r different servers. As pointed in [12], $A_i = 0$ for $i \geqslant k+1$ since for any block of size larger than k, we can reduce the block to k points and the multiset Hall's condition is still satisfied. The following bound is a generalization of the results in [2], [5], [12].

Lemma 6. *If $(X, \mathcal{B})^*$ is an $(n, N, k, m; r)$-MCBC with $r \leqslant k-1$, and A_i for $i \in [k-1]$ is defined as above, then*

$$\sum_{i=r}^{k-1} \binom{m-i}{k-1-i} A_i \leqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}.$$

Proof: Let M_{k-1} be the $\binom{m}{k-1} \times n$ matrix, whose rows are labeled by all the $(k-1)$-subsets of X, and the columns are labeled by the blocks in $\mathcal{B}$ that contain less than k points. The (i, j)-th entry of M_{k-1} is 1 if the j-th block B_j is contained in the i-th $(k-1)$-subset of X, and otherwise it is 0.

Each row in M_{k-1} has at most $\left\lfloor \frac{k-1}{r} \right\rfloor$ ones. In order to verify this property, assume in the contrary that there exist $\left\lfloor \frac{k-1}{r} \right\rfloor + 1$ blocks, and without loss of generality let them be the blocks $B_1, B_2, \ldots, B_{\lfloor \frac{k-1}{r} \rfloor + 1}$, which are all subsets of the same $(k-1)$-subset. Therefore, $|\cup_{i=1}^{\lfloor \frac{k-1}{r} \rfloor + 1} B_i| \leqslant k-1$, and the multiset Hall's condition is not satisfied, since $\lfloor \frac{k-1}{r} \rfloor + 1 = \lceil k/r \rceil$ and $\min\{(\lfloor \frac{k-1}{r} \rfloor + 1)r, k\} = k$. Every column which corresponds to a block of size $i < k$ has exactly $\binom{m-i}{k-1-i}$ ones. Therefore, by counting the number of ones in M_{k-1} by rows and columns separately, we get that $\sum_{i=r}^{k-1} \binom{m-i}{k-1-i} A_i \leqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$. ∎

According to Lemma 6, the next theorem is derived, while its proof follows the one of Lemma 3.2 in [2].

Theorem 7. *Let $r \leqslant k-1$. For any $c \in [r, k-1]$,*

$$N(n, k, m; r) \geqslant nc - \left\lfloor \frac{k-c}{m-k+1} \left[\frac{\left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}}{\binom{m-c}{k-1-c}} - n \right] \right\rfloor.$$

IV. Constructions of MCBCs

In this section we present several constructions of MCBCs. Some of the principles in these constructions use ideas from the constructions of CBCs from [2], [5], [12].

A. A Construction by Replication

Our first construction uses simple replication which is a generalization of the one in [2], [5], [12].

Construction 1 *Let n, k, m, r be positive integers such that $n \geqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$ with $r < k$. We construct an $(n, N, k, m; r)$-MCBC by explicitly constructing its dual set system $(X = [m], \mathcal{B} = \{B_1, \ldots, B_n\})$ as follows:*

1) *The first $\left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$ blocks of $\mathcal{B}$ consist of $\left\lfloor \frac{k-1}{r} \right\rfloor$ copies of all different $(k-1)$-subsets of $[m]$.*
2) *Each remaining block of $\mathcal{B}$ is taken to be any k-subset of $[m]$.*

Thus, the value of N is given by

$$N = kn - \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}.$$

The correctness of this construction is stated in the next theorem.

Theorem 8. *The code $(X, \mathcal{B})^*$ from Construction 1 is an $(n, N, k, m; r)$-MCBC with $n \geqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$, $r < k$ and $N = kn - \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$.*

Before we show that this construction is optimal, let us recall a useful lemma from [2], [5].

Lemma 9. *[[2], [5]] Let $1 \leqslant k \leqslant m$ and $0 \leqslant i \leqslant k-1$. Then $\binom{m-i}{k-1-i} - 1 \geqslant (m-k+1)(k-1-i)$.*

We can now deduce that Construction 1 is optimal.

Corollary 10 *For any $n \geqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$,*

$$N(n, k, m; r) = kn - \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}.$$

Proof: For any $(n, N, k, m; r)$-MCBC, let A_i for $i \in [k]$ be the number of blocks in the dual set system of size i. By Lemma 9, for $i \leqslant k-1$, $\binom{m-i}{k-1-i} \geqslant (m-k+1)(k-1-i) + 1 \geqslant k-i$, then

$$\sum_{i=r}^{k-1} (k-i) A_i \leqslant \sum_{i=r}^{k-1} \binom{m-i}{k-1-i} A_i \leqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1},$$

and the last inequality holds according to Lemma 6. Therefore, we get that

$$\begin{aligned} N &= \sum_{i=r}^{k} i A_i = \sum_{i=r}^{k} (k-(k-i)) A_i = \sum_{i=r}^{k} k A_i - \sum_{i=r}^{k} (k-i) A_i \\ &= kn - \sum_{i=r}^{k-1} (k-i) A_i \geqslant kn - \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}. \end{aligned}$$

Hence, we conclude that $N(n, k, m; r) = kn - \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$ when $n \geqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$, since the codes from Construction 1 achieve this bound. ∎

As a special case when $r = k-1$ we get the following corollary.

Corollary 11

$$N(n, k, m; k-1) = \begin{cases} kn - \binom{m}{k-1} & \text{if } n \geqslant \binom{m}{k-1}, \\ (k-1)n & \text{if } n < \binom{m}{k-1}. \end{cases}$$

B. Constructions Based on Constant Weight Codes

Next, we give constructions based upon constant weight codes. Let (n, d, w)-code denote a binary constant weight code of length n, weight w and minimum Hamming distance d, and let $A(n, d, w)$ denote the maximum number of codewords of an (n, d, w)-code.

Construction 2 *Let $X = [m]$ and $\mathcal{C}$ be an $(m, 2(k-w), w)$-code with n codewords for some $w \in [\max\{r, k/2\}, k-1]$. Let $\mathcal{B} = \{B_1, \ldots, B_n\}$ be the support sets of all the codewords in $\mathcal{C}$.*

The next theorem proves the correctness of this construction.

Theorem 12. *The code $(X, \mathcal{B})^*$ from Construction 2 is an $(n, wn, k, m; r)$-MCBC.*

Proof: We only need to check that $(X, \mathcal{B})$ satisfies the multiset Hall's condition. It is satisfied as the size of each block in $\mathcal{B}$ is $w \geqslant r$ and since the minimum distance of $\mathcal{C}$ is $2(k-w)$, we get that the union of any two blocks in $\mathcal{B}$ is at least k. ∎

For $w = r$ we get the following family of optimal codes.

Corollary 13 *For any $n \leqslant A(m, 2(k-r), r)$, $N(n, k, m; r) = rn$.*

Constant weight codes were used in [2] to construct CBCs. Now, we give a similar construction for MCBCs.

Construction 3 *Let $X = [m]$, $r \leqslant k-2$. Let $\mathcal{C}$ be an $(m, 4, k-2)$-code with α codewords with $\alpha \leqslant A(k, 4, k-2)$. First, let $\mathcal{B}_0$ be a set of $\left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$ blocks, in which each $(k-1)$-subset of $[m]$ appears $\left\lfloor \frac{k-1}{r} \right\rfloor$ times. Let $\mathcal{S}$ consist of the support sets of the codewords in $\mathcal{C}$. Then, for any block in $\mathcal{S}$, add it to $\mathcal{B}_0$, and remove one copy of each of its $m-k+2$ supersets[2] of size $k-1$ in $\mathcal{B}_0$. Let the resulting block set be $\mathcal{B}$.*

Theorem 14. *The code $(X, \mathcal{B})^*$ from Construction 3 is an $(n, N, k, m; r)$-MCBC with*

$$n = \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1} - \alpha(m-k+1) \text{ and } N = n(k-1) - \alpha,$$

where $\alpha \leqslant A(k, 4, k-2)$.

Next, we apply Construction 3 to get a family optimal codes, where we use the bound from $A(n, 4, w) \geqslant \frac{1}{n}\binom{n}{w}$ from [10].

Corollary 15 *For any $\left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1} - (m-k+1) A(m, 4, k-2) \leqslant n \leqslant \left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1}$, $r \leqslant k-2$,*

$$N(n, k, m; r) = n(k-1) - \left\lfloor \frac{\left\lfloor \frac{k-1}{r} \right\rfloor \binom{m}{k-1} - n}{m-k+1} \right\rfloor.$$

[2] For a block $S \in \mathcal{S}$ of size $k-2$, the *supersets* are the $(k-1)$-subsets of $[m]$ that contain S.

C. A Construction for $m = k$

In the following, we give a construction of $(n, N, k, k; r)$-MCBC and determine the value of $N(n, k, k; r)$ for $1 \leqslant r \leqslant k$.

Construction 4 *Let $m = k$ and $X = [k], \mathcal{B} = \{B_1, \ldots, B_n\}$ and $k = \alpha r + \beta$, where $\alpha \geqslant 1$ and $0 \leqslant \beta \leqslant r - 1$ such that the following holds.*

(i) When $\beta = 0$, for any $n \geqslant \alpha$, let $B_i = [(i-1)r + 1, (i-1)r + r]$ for $i \in [\alpha]$, and $B_i = [k]$ for any $i \in [\alpha + 1, n]$.

(ii) When $\beta > 0$, for any $n \geqslant \alpha + r$, let $B_i = [(i-1)r + 1, (i-1)r + r]$ for $i \in [\alpha]$, $B_i = [k] \setminus \{i - \alpha, i - \alpha + r, i - \alpha + 2r, \ldots, i - \alpha + (\alpha - 1)r\}$ for $i \in [\alpha + 1, \alpha + r]$, and $B_i = [k]$ for any $i \in [\alpha + r + 1, n]$.

Theorem 16. *The code $(X, \mathcal{B})^*$ from Construction 4 is an $(n, N, k, k; r)$-MCBC with $N = kn - \left\lfloor \frac{k-1}{r} \right\rfloor k$.*

The next corollary summarizes the construction and results in this section.

Corollary 17 $N(n, k, k; r) = kn - \left\lfloor \frac{k-1}{r} \right\rfloor k$ *if* $r \mid k$, $n \geqslant \frac{k}{r}$ *or* $r \nmid k$, $n \geqslant \lfloor \frac{k}{r} \rfloor + r$.

D. A Construction from Steiner Systems

In the following we construct a class of MCBCs based upon Steiner systems, which is a generalization of Example 2.

A *Steiner system* $S(2, \ell, m)$ is a set system $(X, \mathcal{B})$, where X is a set of m points, $\mathcal{B}$ is a collection of ℓ-subsets (blocks) of X, such that each pair of points in X occurs together in exactly one block of $\mathcal{B}$. For the existence of Steiner systems, we refer the reader to [9].

Theorem 18. *Let $(X, \mathcal{B})$ be an $S(2, \ell, m)$ with $m > \ell$. Then $(X, \mathcal{B})^*$ is a $(|\mathcal{B}|, \ell|\mathcal{B}|, k, m; r)$-MCBC for any $\lfloor \frac{\ell}{2} \rfloor + 1 \leqslant r \leqslant \ell$ and $k \leqslant (\ell - r + 1)(2r - 1)$.*

An *affine plane* of order q is an $S(2, q, q^2)$. It has q^2 points and $q^2 + q$ blocks. It is well known that an affine plane exists for any prime power q [9]. The next result of CBCs based upon affine planes was given in [15].

Theorem 19. *[15] Let q be a prime power and $(X, \mathcal{B})$ be an affine plane of order q. Then $(X, \mathcal{B})^*$ is an optimal uniform $(q^2 + q, q^3 + q^2, q^2, q^2)$-CBC.*

The code in Theorem 19 is also an optimal CBC. However, note that it is a different code from the optimal (n, N, k, k)-CBC in [12] which is constructed as follows: Let $X = [k]$, and $\mathcal{B} = \{B_1, \ldots, B_n\}$, which are given by $B_i = \{i\}$ for $i \in [k]$, and $B_i = [k]$ for $i \in [k + 1, n]$. By Theorem 18, we can see the code in Theorem 19 is also $(q^2 + q, q^3 + q^2, k, q^2; r)$-MCBCs for different pair-values of k and r.

Corollary 20 *Let q be a prime power. Then there exists a $(q^2 + q, q^3 + q^2, k, q^2; r)$-MCBC for any $\lfloor \frac{q}{2} \rfloor + 1 \leqslant r \leqslant q$ and $k \leqslant (q - r + 1)(2r - 1)$.*

When $r = q$, we also receive an optimal $(q^2 + q, q^3 + q^2, 2q - 1, q^2; q)$-MCBC, since it reaches the bound in Lemma 5 (i) with total storage $N = rn = q(q^2 + q) = q^3 + q^2$. Note that this code could also be obtained by Construction 2 using $(q^2, 2q - 2, q)$-codes. The existence of $(q^2, 2q - 2, q)$-codes follows from affine planes as follows. For any block $B \in \mathcal{B}$, we get a codeword u of length q^2 in which the value of each coordinate u_i for $i \in [q^2]$ is 1 if and only if $i \in B$. Since any two blocks intersect in at most one points, the distance between every two distinct codewords is at least $2(q - 1)$. Finally, we note that it is possible to improve the value of k when $r \leqslant \lfloor \frac{q}{2} \rfloor$, and the optimality of this code construction for other values of r is left open for future research.

V. Conclusion

In this paper, we generalized combinatorial batch codes to multiset combinatorial batch codes. Several bounds and constructions of optimal codes were obtained.

Acknowledgments

The authors would like to thank Prof. Tuvi Etzion for valuable discussions. The work of Hui Zhang is supported in part at the Technion by a fellowship of the Israel Council of Higher Education.

References

[1] N. Balachandran and S. Bhattacharya, "On an extremal hypergraph problem related to combinatorial batch codes," *Discret. Appl. Math.*, vol. 162, pp. 373–380, 2014.

[2] S. Bhattacharya, S. Ruj, and B. Roy, "Combinatorial batch codes: a lower bound and optimal constructions," *Adv. Math. Commun.*, vol. 6, pp. 165–174, 2012.

[3] R. A. Brualdi, K. P. Kiernan, S. A. Meyer, M. W. Schroeder, "Combinatorial batch codes and transversal matroids," *Adv. Math. Commun.*, vol. 4, pp. 419–431, 2010.

[4] C. Bujtás and Z. Tuza, "Combinatorial batch codes: extremal problems under Hall-type conditions," *Electr. Notes in Discrete Math.*, vol. 38, pp. 201–206, 2011.

[5] C. Bujtás and Z. Tuza, "Optimal batch codes: many items or low retrieval requirement," *Adv. Math. Commun.*, vol. 5, pp. 529–541, 2011.

[6] C. Bujtás and Z. Tuza, "Optimal combinatorial batch codes derived from dual systems," *Miskolc Math. Notes*, vol. 12, no. 1, pp. 11–23, 2011.

[7] C. Bujtás and Z. Tuza, "Relaxations of Hall's condition: optimal batch codes with multiple queries," *Appl. Anal. Discret. Math.*, vol. 6, no. 1, pp. 72–81, 2012.

[8] C. Bujtás and Z. Tuza, "Turán numbers and batch codes," *Discret. Appl. Math.*, vol. 186, pp. 45–55, 2015.

[9] C. J. Colbourn and R. Mathon, *Steiner systems*, Handbook of Combinatorial Designs (C. J. Colbourn and J. H. Dinitz, eds.), Chapman & Hall/CRC, Boca Raton, 2nd ed., pp. 102–110, 2007.

[10] R. L. Graham and N. J. A. Sloane, "Lower bounds for constant weight codes," *IEEE Trans. Inform. Theory*, vol. 26, pp. 37–43, 1980.

[11] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in *Proc. of the 36-sixth Annual ACM Symposium on Theory of Computing STOC '04*, pp. 262–271, 2004.

[12] M. B. Paterson, D. R. Stinson, and R. Wei, "Combinatorial batch codes," *Adv. Math. Commun.*, vol. 3, pp. 13–27, 2009.

[13] A. S. Rawat, D. S. Papailiopoulos, A. G. Dimakis, and S. Vishwanath,, "Locality and availability in distributed storage," *IEEE Trans. Inform. Theory*, vol. 62, pp. 4481–4493, 2016.

[14] N. Silberstein, "Fractional repetition and erasure batch codes," in *4th International Castle Meeting on Coding Theory and Applications*, Palmela, Portugal, Sep. 2014.

[15] N. Silberstein and A. Gál, "Optimal combinatorial batch codes based on block designs," *Des. Codes Cryptogr.*, pp. 1–16, 2014.

[16] A. Zeh and E. Yaakobi, "Bounds and constructions of codes with multiple localities," in *Proc. IEEE Intl. Symp. Inform. Theory*, Barcelona, Spain, pp. 640–644, Jul. 2016.